DOI: 10.16210/j.cnki.1007-7561.2024.03.023
侯晓龙, 杨卫东, 李磊, 等. 基于骨架序列多算法的粮仓作业人员异常行为视频识别[J]. 粮油食品科技, 2024, 32(3): 201-210.
HOU X L, YANG W D, LI L, et al. The video recognition of abnormal behavior of granary workers based on skeleton sequence multi-algorithm fusion[J]. Science and Technology of Cereals, Oils and Foods, 2024, 32(3): 201-210.

基于骨架序列多算法的粮仓作业人员异常行为视频识别

侯晓龙[1]，杨卫东[1]✉，李　磊[1]，于俊伟[2]，许启铿[3]

（1. 河南工业大学 信息科学与工程学院，河南 郑州 450001；
2. 河南工业大学 人工智能与大数据学院，河南 郑州 450001；
3. 河南工业大学 土木工程学院，河南 郑州 450001）

摘　要：粮仓是保障粮食储藏安全的重要设施。粮仓为封闭大空间，仓内光照昏暗、空气流通差，熏蒸、气调等作业增加了人员安全隐患，通过仓内安防视频对作业人员的异常行为进行识别与分析，是作业人员安全操作的一项重要技术保障。提出了一种基于骨架序列多算法的粮仓内作业人员异常行为的视频识别算法。首先，利用 YOLOv3tiny 模型对人体进行快速检测，结合 Sort 对多目标进行运动轨迹跟踪，通过 AlphaPose 模型提取人体骨架坐标序列及权重信息；进而，根据人体骨架自然连接节点构成的实际空间图（RSG）和虚拟人体的重心与头、手、脚互连构建的虚拟空间图（VSG），基于人体动力学重心与手脚互动的平衡性，提取仓内作业人员异常行为的空间特征和串联时间卷积（TC）的时空特征；最后，提出了虚实结合的时空图卷积网络（VR-STGCN）仓内作业人员的异常行为视频识别算法。同时自建了混合数据集，并将 VR-STGCN 与 SSD、PCANet、Two-StreamCNN、STGCN 等四种算法进行了对比实验与分析。结果表明：VR-STGCN 各项指标均优于其他四种算法；VR-STGCN 能够在光线不足、多目标、远距离等复杂环境下准确地识别出仓内人员的跌倒、爬行、躺平等异常行为，识别准确率达到 97.7%，处理速度为 18.67 fps，能够实时分析作业人员异常行为。研究成果为复杂环境下粮仓作业人员的安全保障提供了一种全新高效的技术。

关键词：时空图卷积；异常行为识别；人体动力学；粮仓作业安全

中图分类号：TP391　　**文献标识码：**A　　**文章编号：**1007-7561(2024)03-0201-10

网络首发时间：2024-05-10 08:35:03

网络首发地址：https://link.cnki.net/urlid/11.3863.ts.20240508.2156.004

收稿日期：2024-01-08

基金项目：河南省重大公益专项（201300210100）；河南省杰出青年基金（222300420004）；2021 年度河南省重点研发与推广专项（212102210152）

Supported by: Major Public Project of Henan Province (No. 201300210100); Outstanding Youth Fund of Henan Province (No. 222300420004); Key Research and Development and Promotion Projects of Henan Province in 2021 (No. 212102210152)

作者简介：侯晓龙，男，1989 年出生，硕士，工程师，研究方向为计算机视觉、大数据。E-mail: houxiaolong@haut.edu.cn

通讯作者：杨卫东，男，1977 年出生，博士，教授，研究方向为物联网系统及安全、智能网联汽车安全、计算机视觉等。E-mail: yangweidong@haut.edu.cn

The Video Recognition of Abnormal Behavior of Granary Workers Based on Skeleton Sequence Multi-algorithm Fusion

HOU Xiao-long[1], YANG Wei-dong[1]✉, LI Lei[1],YU Jun-wei[2], XU Qi-keng[3]

(1. College of Information Science and Engineering, Henan University of Technology, Zhengzhou, Henan 450001, China; 2. School of Artificial intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan 450001, China; 3. College of civil engineering and architecture, Henan University of Technology, Zhengzhou, Henan 450001, China)

Abstract: Granary is an important facility to ensure the safety of grain storage. The granary is a large closed space with dim lighting and poor air circulation. Operations such as fumigation and air conditioning increase personnel safety risks. The identification and analysis of abnormal behaviors of workers through security videos in the granary has become a key safe operations for workers as an important technical guarantee. This paper proposed a video recognition algorithm for abnormal behavior of workers in a granary based on a skeleton sequence multi-algorithm. First, the YOLOv3tiny model was used to quickly detect the human body, combined with Sort to track the motion trajectories of multiple targets, and the human skeleton coordinate sequence and weight information were extracted through the AlphaPose model. Then, based on the real spacial graph (RSG) composed of natural connection nodes of the human skeleton and virtual spacial graph (VSG) constructed by interconnecting the center of gravity of the virtual human body with the head, hands, and feet, the bin was extracted based on the balance of the interaction between the center of gravity of the human body dynamics and the hands and feet. Spatial characteristics of abnormal behavior of internal workers and spatiotemporal characteristics of concatenated temporal convolution (TC). Finally, a virtual-real combining spatial temporal graph convolution network (VR-STGCN) video recognition algorithm for abnormal behavior of granary workers was proposed. At the same time, a hybrid dataset was built, and comparative experiments and analysis were conducted between VR-STGCN and four algorithms such as SSD, PCANet, Two-StreamCNN, and STGCN. The results showed that all indicators of VR-STGCN were better than those of the other four algorithms. VR-STGCN can accurately identify abnormal behaviors such as falling, crawling, and lying down of people in the granary in complex environments such as insufficient light, multiple targets, and long distances. The recognition accuracy reached 97.7%, and the processing speed was 18.67fps, which can analyze the abnormal behavior of workers in real time. The research results could provide a new and efficient technology for the safety of granary workers in complex environments.

Key words: spatial-temporal graph convolution; identification of abnormal behavior; action estimation; grain storehouse operation safety

目前，我国粮食年产量约 6.6 万 t，做好粮食储备，粮仓作业人员安全意义重大。粮仓是保障粮食储藏安全的重要设施。粮仓为封闭大空间，仓内光照昏暗、空气流通差，熏蒸、气调等作业增加了人员安全隐患。根据国家粮食购销领域监管信息化规范，要求中央和地方政府事权粮食的承储库点的视频监控系统互联互通；同时要求建设视频图像分析边缘管控，实现仓内监控实时智能分析，支撑粮食购销信息化监管水平。通过仓内安防视频对作业人员的异常行为进行识别与分析，成为了作业人员安全操作的一项重要技术保障，为粮库安全生产、粮食购销监管提供了较好技术支撑。

早期，对于人员的行为识别主要采用手动特征提取，包含全局特征提取和局部特征提取，并通过直接分类和基于模型分类的方式进行行为识

别。其中全局特征提取对视频帧整体进行处理，运用目标跟踪算法或者运动背景减除定位出人物，之后把选定处理的目标区域通过编码方式提取特征。而局部特征的提取分为基于时空兴趣点和基于轨迹追随两种方式，其中，基于时空兴趣点的特征提取方法侧重于识别和提取图像中的特定空间位置和时间点上的特征，而基于轨迹追随的特征提取方法则更加注重于时空关系的建模和跟踪。

然而，手动特征提取是一项繁琐和耗时的任务，并且易受背景抖动、光照变化及图像噪声等影响很难捕捉到复杂的行为模式。近年来，随着深度学习方法的兴起，可以实现自动特征学习和行为识别。视频动作识别的表现形式是分类任务（video-classfication），本质技术是时空特征学习（spatio-temporal feature learning）[1]，基于深度学习的行为识别方法以端到端的方式从数据中学习特征，目前主流的前沿视频动作识别算法包含基于 CNN、vision-transformer、self-supervised、multimodal 的方法。CNN 方法主要分为三方面，一是在 2D 卷积基础上，引入 temproal 建模能力，如 two-stream[2]，TSN[3]；二是基于 3D 卷积的改进，如 I3D[4]、P3D[5]、R(2+1)D[6]；三是更强时空建模能力的探索，如 non-local[7]、slowfast[8]。基于 Vision- Transformer[9] 的方法主要模型有 TimeSformer[10]、Pitch Shift Transformer[11]、Video Swin Transformer[12-13]，对输入的连续采样的多帧图像（8 帧、16 帧、32 帧）的数据，学习视频的时域（Temporal）和空间（Spatial）特征。Vision-Transformer 首先在图像分类领域取得成功，性能超过 CNN。基于 self- supervised 的方法采用 Video Masked Autoencoders（MAE）[14-16]自监督思想。基于 multimodal[17-19]的方法通过融合来自不同感知模态的特征，能够从多个角度和多个维度对动作进行建模，提高了动作识别的准确性和鲁棒性。然而，以上深度学习算法模型参数计算量大，硬件算力要求高，实用成本高，时间规模与实时性差，无法满足库区仓内作业人员异常行为的实时识别。

对于库区仓内特殊环境下的作业人员异常行为的实时准确识别，本文提出了一种基于骨架序列多算法的粮仓作业人员异常行为视频识别算法。首先，采用 Yolov3tiny 模型对人体进行快速检测，结合 SORT 对多目标进行运动轨迹跟踪；然后，利用 AlphaPose 模型提取人体骨架序列；最后通过 VR-STGCN 网络对仓内作业人员的异常行为进行识别。充分利用人体骨关节的活动轨迹与时空关系对目标行为进行实时的准确识别。骨架序列多算法的粮仓内作业人员异常行为视频识别算法流程如图 1。

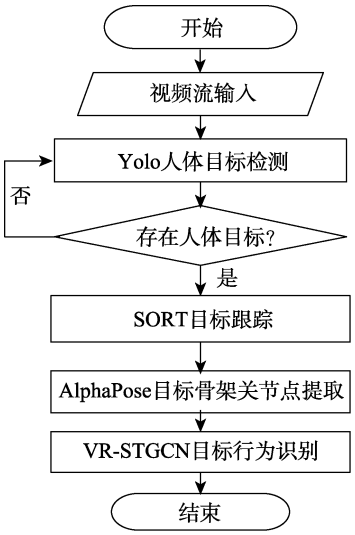


图 1 骨架序列多算法流程图

Fig.1 Flow chart of skeleton sequence fusion algorithm

本文根据人体动力学，重心与手脚互动的平衡性，符合人体动作识别的行为逻辑，充分提取时空特征，识别异常行为。本文主要贡献如下：

（1）提出根据人体动力学平衡性，重心与头、手、脚互连，挖掘重心轴与重力脚的关系，创建虚拟空间图（VSG），用于二维数据骨架关节点识别人体异常行为。

（2）提出利用自然连接骨架关节点所构成的真实空间图（RSG），重心点与头、手、脚互连的虚拟空间图（VSG），对邻接矩阵进行改造，扩展多标签子集，同时加大对重心点置信度的权重，有效避免邻接矩阵特征提取同质化和 GCN 过平滑问题。提出 VR-STGCN 虚实时空图卷积网络，增大卷积核，增强网络感受野，提高获取全局特征的能力。

（3）利用骨架序列多算法融合对粮仓作业人

员进行异常行为识别，通过仓内安防视频设备对仓内作业人员提供安全、实时预警，保障国家仓储人员作业安全。

1 人体目标检测

准确的人体目标检测是行为识别的关键，YOLO(You only look once)是一种基于卷积神经网络的实时目标检测算法。它的主要特点是在一次前向传递中同时进行物体检测和分类。YOLOv3tiny 相比于其他版本的 YOLO 具有更轻量级的网络结构。它使用了仅有 13 个卷积层的特殊网络结构，以减少参数量和计算复杂度，速度快，适合应用于实时场景[20]。

YOLOv3tiny 网络使用 Darknet-53 作为主干网络，以 416*416*3 的图像作为输入，通过将待检图片划分成为 S*S 个网格，有 2 个输出分支进行多尺度预测。每个网格会产生 3 个 anchorbox，每个 anchorbox 包含宽高、中心坐标、置信度及每个类别的预测值，根据超参置信度阈值，筛除得分低的预测框，对剩余预测框执行非极大值抑制（Non-Maxinum Suppression,NMS），得到最终预测框。置信度 C 表示如下：

$$C = pre(object) * \mathrm{IoU}_{\mathrm{pred}}^{\mathrm{truth}} \quad \text{式（1）}$$

其中，IoU 是预测边框与实际边框的重合度，pre（object）为网格中分类对象的概率。

2 目标跟踪

SORT（Simple Online and Realtime Tracking）是一种用于多目标跟踪的算法。首先，提供视频的原始帧，运行 faster-rcnn[21]目标检测器，识别出视频帧中的目标物体，以获取目标检测框。提取目标检测框中目标的外观和其他属性特征并编码。

本文修改 SORT 目标检测算法，由于 faster-rcnn 目标检测器是一个 two-stage 的检测算法，也就是把检测问题分成了两个阶段，第一个阶段是生成候选区域，第二个阶段是对候选区域位置进行调整以及分类。这种方法速度很慢，很难达到实时检测的效果。采用端到端 one-stage 算法 yolo 作为目标检测器，并利用卡尔曼滤波算法和匈牙利算法，极大提高了多目标跟踪的速度。

SORT 算法通过计算前后两帧目标框的交并比（IOU）来构建相似度矩阵。这种方法避免了复杂的相似度计算，使得 SORT 算法具有快速的计算速度。图 2 展示 SORT 核心算法流程。

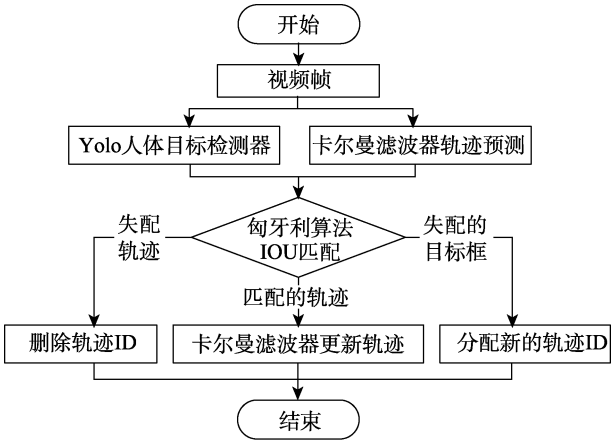


图 2 SORT 核心算法流程图

Fig.2 SORT core algorithm flowchart

其中，通过 Yolo 目标检测器获取目标框（Detections），同时使用卡尔曼滤波器对当前帧的轨迹进行预测和更新（Tracks）。然后，通过计算目标框和轨迹之间的交并比（IOU）来进行匹配。根据匹配结果，可以将得到的结果分为 3 类：（1）失配的轨迹：这部分轨迹与任何目标框均无法匹配，被视为失配。如果连续 30 次失配，该轨迹的目标 ID 将从当前帧中删除。（2）失配的目标框：这部分目标框无法与任何轨迹匹配，因此需要为这些目标框分配新的轨迹。（3）匹配的轨迹：这部分轨迹与目标框成功匹配。匹配后，卡尔曼滤波器可以利用轨迹的状态来预测下一帧的目标框状态。卡尔曼滤波器的更新过程会使用观测值（即匹配上的轨迹）和估计值来更新所有轨迹的状态。

SORT 算法通过目标检测和卡尔曼滤波的预测与更新过程，将目标框和轨迹进行匹配，并根据匹配结果进行轨迹的更新和新轨迹的创建。这样可以实现对多目标的跟踪和预测。

3 提取骨架关节点

AlphaPose[22]姿态识别框架采用自顶向下的方法，通过 yolo 目标检测器得到人体检测框，Alphapose 提出每一个检测框中检测人体关键点，

连接成一个人形，实现区域多人姿态检测框架。该框架主要由 3 个部分组成，对称空间变换网络（Symmetric Spatial Transformer Network，SSTN）、姿态引导的样本生成器（Pose-Guided Proposals Generator，PGPG）和姿态非极大值抑制器(Parametric Pose Non Maximum-Suppression，PNMS)。

SSTN 是对称空间变换网络，由 STN 和 SDTN 两部分组成。STN 负责接收人体候选框，SDTN 则生成候选姿态。通过这两个部分的组合，可以实现对人体区域定位和姿态检测的精确性。

PNMS 是一种参数化非极大值抑制法，通过定义姿态距离度量来衡量姿态的相似度。当姿态距离度量小于设定的阈值时，多余的姿态估计将被过滤掉，从而提高骨骼关键点检测的准确性。

PGPG 是姿态引导的样本生成器，它是 SPPE 部分的一部分。PGPG 可以对已有的数据进行增强扩充，帮助 SSTN 适应不完美的人体区域定位和姿态检测。通过 PGPG 生成各种姿态的图片，用于训练过程。

4 行为识别

本文采用虚实结合的时空图卷积网络（Virtual-Real Combining Spatial Temporal Graph Convolution Network，VR-STGCN），利用 Alphapose 算法获得每一帧图像人体目标 14 个骨架关节点的二维坐标和置信度信息，根据空间和时间关节点的变化提取运动特征，进行行为识别。针对时空图卷积网络（STGCN）在行为识别方面存在指利用二维坐标关节点数据而忽略人体在运动过程中，自身重心在打破人体平衡，产生异常行为的作用。

4.1 骨架序列时空图设计

利用人体有效的骨架关节点二维坐标与置信度数据，根据人体关节点的自然连接建立真实空间图（RSG），同时，根据重心轴与重心脚的人体动力学平衡关系，建立重心与头、手、脚互连的虚拟空间图（VSG）。在时间序列采用各关节点沿时间轴互连构成各关节点的时间图（TGN），采用 VR-STGCN 完成人体骨架学列的时空图卷积。图 4 展示了重心轴与重心脚的人体动力学平衡关系。

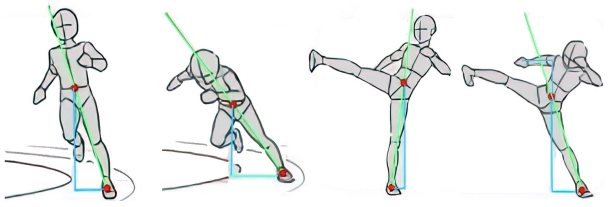
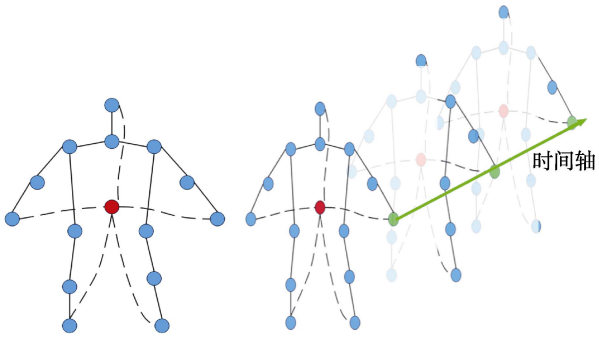


图 3 重心轴与重心脚的人体动力学平衡关系

Fig.3 The human dynamic balance relationship between the center of gravity axis and the center of gravity foot

定义 1，骨架序列图结构为 $G=\{V_{vr},E_{vrt}\}$ 表示，其中，V_{vr} 代表两类骨架关节点，一是真实存在的骨架关节点 V_r，一是虚拟的重心点 V_v。E_{vrt} 代表 3 类连接边，E_r 是关节点的自然连接边，E_v 是重心点与头、手脚互连的虚拟边，E_t 是单一关节点沿时间轴互连边，如图 4 骨架序列图结构所示。



(a) E_t自然连接边、E_v虚拟边　　(b) E_i沿时间轴互连边

图 4 骨架序列图结构

Fig.4 Skeleton sequence graph structure

4.2 虚实时空图卷积设计

虚实时空图卷积设计包含空间卷积设计、时间卷积设计和 VRSTGCN 卷积设计 3 部分。

4.2.1 空间卷积设计

本文采用多标签方法对空间图建模实现卷积。

定义 2，给定骨架序列图 G，自然连接边的邻域表示为，$B_r(v_i)=\{v_j \mid (v_j,v_i)\in E_r\},(\forall_{vi}\in V_r)$，虚拟连接边的邻域表示为 $B_v(v_i)=\{v_j \mid (v_j,v_i)\in E_v\},(\forall_{v_i}\in V_v)$，存在映射函数 $l_i(B(v_i))\rightarrow\{0,1,2,3,\cdots,K\}$。这种通过扩展邻域分类的方式称为多标签技术。利用多标签方法的图卷积通用公式（2）为：

$$f(v_i)=\sum_{v_j\in B(v_i)}\frac{1}{Z_i(v_j)}f_{in}(v_j)\cdot W(l_i(v_j)) \quad 式（2）$$

其中，$f_{in}(v_j)$ 是关节点 v_i 的邻域节点的输入特征，$W(l_i(v_j))$ 是关节点 v_i 通过多标签选择的邻接矩阵的权重函数，$Z_i(v_j)$ 是对应邻接矩阵的度正则化表示，以均衡不同子集的贡献。

定义 3，真实空间图（RGN）的多标签方法是根据关节点 v_i 与中心点（人体左右肩中间位置）的距离，选择为自身根节点、离心点、向心点 3 个不同子集。标签数集为 $\{0,1,2\}$，数学公式（3）为：

$$l_i(v_i)=\begin{cases}0, v_j\in V_r,(v_j,v_i)\in E_r, d_{(v_j,c)}=d_{(v_i,c)}\\ 1, v_j\in V_r,(v_j,v_i)\in E_r, d_{(v_j,c)}<d_{(v_i,c)}\\ 2, v_j\in V_r,(v_j,v_i)\in E_r, d_{(v_j,c)}>d_{(v_i,c)}\end{cases} \quad 式(3)$$

其中，c 是中心点（人体左右肩中间位置）。

定义 4，虚拟空间图的标签方法是关节点 v_i 与重心点 g 直连。如公式（4）所示：

$$l_i(v_i)=3, v_j\in V_v,(v_j,v_i)\in E_v \quad 式(4)$$

其中重心 g 位置为人体左右肩中心点与左右胯骨中心点连线，距离左右肩中心点 3/4 处。

扩展后的多标签对应 4 个子集表示如图 5 所示：

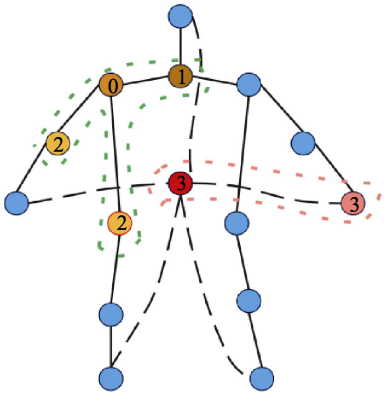


图 5 扩展多标签子集

Fig.5 Extended multi-label subset

定义 5，根据定义 3 和定义 4 所构成的虚实空间图，利用定义 2 的多标签技术对邻接矩阵进行扩展，构成虚实结合的时空图卷积网络模型的邻接矩阵。如公式（5）所示：

$$A=A^r+A^c+A^f+A^v=\sum_{q\in Q}A^q \quad 式(5)$$

其中，$Q=\{r,c,f,v\}$，r 是关节点根节点，c 是向心点，f 是离心点，v 是虚拟重心与头、手、脚互连节点。

由此，构建虚实结合图的卷积公式（6）所示：

$$f_{out1}=\sum_{q\in Q}M^q\otimes((D^q)^{-1/2}A^q(D^q)^{-1/2})F_{in}\cdot W^q \quad 式(6)$$

其中，D 为邻接矩阵 A 子集的度，$(D^q)^{-1/2}A^q(D^q)^{-1/2}$ 实现对邻接矩阵 A 子集的归一化，M 是邻接矩阵 A 子集的边的权重，$\otimes$ 哈达玛内积是两个相同尺寸的矩阵对应位置元素的乘积。W 是关节点的权重函数，虚拟重心的初始权重不小于 1。虚实结合图的卷积的简化图如图 6 所示：

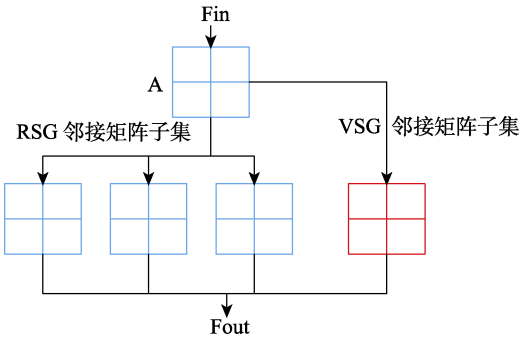


图 6 虚实结合图卷积简化图

Fig.6 Simplified graph of virtual real combining graph convolution

4.2.2 时间卷积设计

人体骨架各关节点沿时间轴构成各节点的时间序列，该时间序列构成各节点的时间图，通过二维时间卷积网络 TC 进行特征提取，假设卷积核大小定义为 Γ，则该节点 t 时刻的邻域 $B(t)=\{p\mid p-t|<\lfloor\Gamma/2\rfloor\}$，虚拟重心点与真实关节点沿时间轴的时间卷积 TC 表示为公式（7）：

$$f_{out2}=\sum_{p\in B(t)}F_{in}(p)W^p \quad 式(7)$$

其中，F_{in} 是节点 p 的输入特征，W 是权重函数。

4.2.3 VR-STGCN 卷积设计

虚实结合的时空图卷积采用对空间卷积和时间卷积特征提取串联方式。对输入视频帧中人体目标的二维骨架关节点数据的平滑和标准化预处理，同时引入残差机制，首先对虚拟空间图与真实空间图特征的融合，将融合特征输入时间卷积提取时空特征，实现虚实结合的时空图卷积。公式（8）如下：

$$f_{out}=f_{out2}(f_{out1}(G)) \quad \text{式（8）}$$

其中，f_{out2} 是公式 6，f_{out1} 是公式 5，G 是定义 1 中的骨架序列图结构。

虚实结合的时空图卷积如图 7 所示：

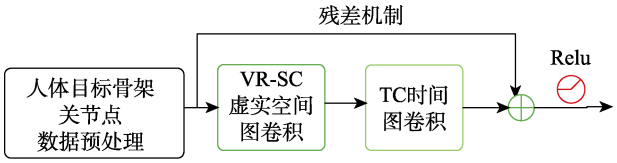


图 7 虚实结合的时空图卷积简化图

Fig.7 Simplified graph of virtual-real spacial temporal graph convolution

4.3 虚实结合的时空图卷积网络模型

虚实结合的时空图卷积网络模型通过 10 层 VR-STGCN 卷积模块来提取更深层次的特征，最后，通过池化、全连接、SoftMax 分类器计算每个动作类别的概率，并输出概率最大的动作类别作为最终的预测结果。损失函数采用交叉熵函数，用于度量模型输出的概率分布与真实标签之间的差异。通过最小化交叉熵损失函数，可以使模型更好地学习分类任务，其计算表达式为公式（9）：

$$L=-\sum_{c=1}^{C} y_c \log_2(p_c) \quad \text{式（9）}$$

其中，c 表示类别；C 表示类别数量；y_c 表示指示变量（0 或 1），当该类别和样本类别一致时为 1，反之为 0；p_c 表示类别 c 的预测概率。

利用虚实结合的时空图卷积网络实现骨架序列多算法融合的粮仓作业人员异常行为识别的简图如图 8 所示。

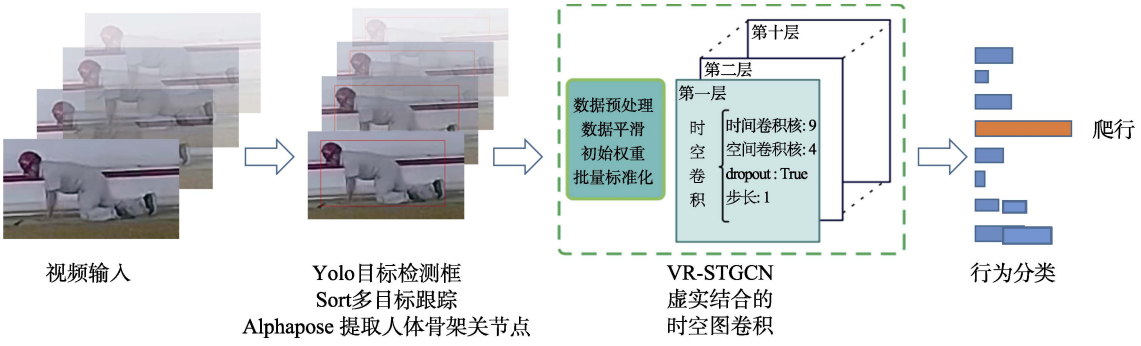


图 8 虚实结合的时空图卷积网络模型简化图

Fig.8 Simplified graph of virtual real combining spatial-temporal graph convolution network

5 实验结果与模型分析

本文实验采用硬件配置为 inter i5-7500 处理器，内存 RAM 为 16G，显卡为 NVIDIA GEFORCE RTX 3060。

5.1 数据集

采用自建混合数据集，包含 100 段 le2ifalldataset 中办公室和教室日常公共数据集和 40 段粮仓内作业人员活动数据集，其中 le2i 格式为 325*280 像素，25 帧，大小 10 s 左右；粮仓内作业人员活动数据集 1 080*1 920 像素，25 帧，大小 10 s 左右，仓内人员活动离摄像机距离范围为 1~25 m。视频包含站立、走动、坐、躺、站起、坐下、跌倒、爬行等 8 类动作。数据集中 8 类动作共计约 5 万帧的样本分布情况如图 9。

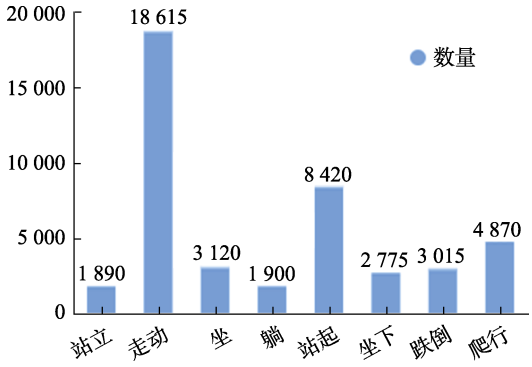


图 9 自建混合数据集样本分布

Fig.9 Sample distribution of self-built mixed dataset

5.2 实验结果

骨架序列多算法采用 yolo 目标检测对人员目标进行检测框的标定，Sort 和 alphapose 根据目标检测框进行目标跟踪和 14 个骨架关节点的提取，

最后输入 VR-STGCN 进行行为识别。

采用自建混合数据集，训练集占样本的 75%，测试集占 25%，模型的 epochs 为 30，batch size 为 32，采用一种自适应学习率 Adadelta 优化器来更新权重。模型验证结果准确率为 98%，处理速度达到 18.67 fps，如图 10。

本文采用混淆矩阵、准确率、召回率、精确度和 F1 分数评估模型性能的指标。

表 1 展示了通过测试获得虚实混合时空图卷积网络模型在混合数据集上的混淆矩阵。

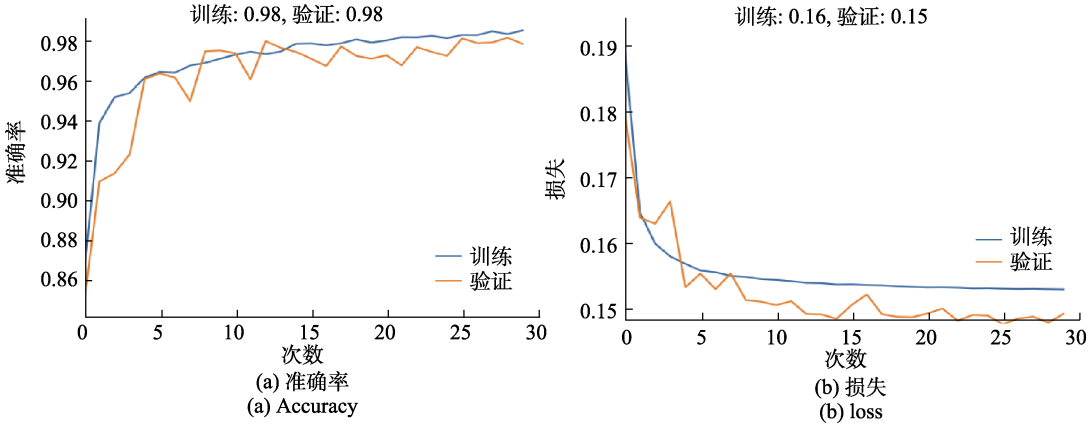


图 10 自建混合数据集上准确率及损失

Fig.10 Accuracy and loss on self-built mixed dataset

表 1 混合数据集混淆矩阵

Table 1 Mixed dataset confusion matrix

	混合数据集混淆矩阵							
	站立	走动	坐	躺	站起	坐下	跌倒	爬行
站立	0.973	0.009	0.000	0.000	0.000	0.000	0.018	0.000
走动	0.008	0.977	0.000	0.000	0.003	0.006	0.006	0.000
坐	0.000	0.011	0.987	0.000	0.000	0.002	0.000	0.000
躺	0.000	0.000	0.000	0.971	0.007	0.000	0.021	0.000
站起	0.000	0.016	0.005	0.000	0.969	0.002	0.008	0.000
坐下	0.008	0.008	0.030	0.000	0.004	0.951	0.000	0.000
跌倒	0.000	0.000	0.000	0.005	0.000	0.000	0.984	0.010
爬行	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.998

通过实验数据分析，虚实结合时空图卷积骨架序列融合算法在粮仓内对作业人员的行为识别精确度为 97.7%、召回率为 97.67%、F1 值为 97.68、平均损失为 0.1524、准确率为 97.6%。

5.3 实验对比

本文在动作识别阶段与 SSD[23]、PCANet[24]、Two-Stream[25]、STGCN[26] 4 种模型在准确率、召回率、F1 值进行对比。对比如表 3 所示

表 3 中的数据表明，VR-STGCN 在精确度、召回率和 F1 值等评估指标上均优于其他 4 种算法。此外，与 STGCN 相比，VR-STGCN 的预测精确度提高了 0.7%。这表明虚实结合的时空图卷积方法可以提高模型的特征提取能力，从而优化模型的性能。

表 2 骨架序列融合算法性能指标

Table 2 Performance indicators of skeleton sequence fusion algorithm

	模型性能指标		
标签	精确度/%	召回率/%	F1 值
站立	98.41	97.32	97.86
走动	95.75	97.75	96.74
坐	96.55	98.72	97.62
躺	99.47	97.14	98.29
站起	98.57	96.93	97.74
坐下	99.02	95.06	97
跌倒	94.89	98.44	96.63
爬行	98.97	99.8	99.48
精确度=97.7%，召回率=97.67%，F1 值=97.68，平均损失：0.152 4，准确率：97.7%			

表 3 不同算法检测结果对比

Table 3 Comparison of detection results using different algorithms

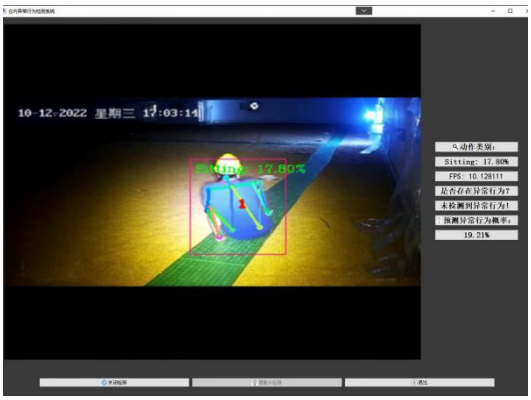
不同算法检测结果			
算法	精确度/%	召回率/%	F1 值
SSD	89.5	87.00	90.00
PCANet	91.8	90.00	91.00
Two-Stream CNN	94.2	92.00	94.00
STGCN	97.0	95.00	96.10
VR-STGCN	97.7	97.67	97.68

5.4 检测平台实现

图 11 展示使用 pyqt5 搭建异常行为检测平台，对实时视频和 MP4 视频流进行实时分析。

(a) 仓内作业人员仓内跌倒检测

(a) Fall Down posture detection of workers in granary



(b) 仓内作业人员坐姿检测

(b) Sitting posture detection of workers in granary

图 11 仓内异常行为检测平台展示图

Fig.11 Display diagram of abnormal behavior detection platform in granary

6 结语

通过多算法提取人体骨架关节点序列，利用虚实结合的时空图卷积网络（VR-STGCN）在光线不足、多目标远距离等复杂的环境下，提取粮仓内作业人员异常行为的时空特征，对跌倒、爬行、躺平等异常行为识别准确率 97.7%，处理速度达到 18.67 fps，能够满足仓内作业人员异常行为的实时分析，进而预防安全事件的发生。该研究成果为复杂环境下粮仓作业人员的安全保障提供了一种全新高效的技术。

参考文献：

[1] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks[C]// Proceedings of the IEEE international conference on computer vision, 2015: 4489-4497.

[2] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in neural information processing systems, 2014, 27.

[3] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]// European conference on computer vision. Springer, Cham, 2016: 20-36.

[4] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6299-6308.

[5] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3d residual networks[C]//proceedings of the IEEE International Conference on Computer Vision, 2017: 5533-5541.

[6] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018: 6450-6459.

[7] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 7794-7803.

[8] FEICHTENHOFER C, FAN H, MALIK J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision, 2019: 6202-6211.

[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[10] BERTASIUS G, WANG H, TORRESANI L. Is space-time attention all you need for video understanding?[C]//ICML. 2021, 2(3): 4.

[11] XIANG W, LI C, WANG B, et al. Spatiotemporal self-attention modeling with temporal patch shift for action recognition[C]// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 627-644.

[12] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the

IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[13] LIU Z, NING J, CAO Y, et al. Video swin transformer[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 3202-3211.

[14] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16000-16009.

[15] TONG Z, SONG Y, WANG J, et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training[J]. Advances in neural information processing systems, 2022, 35: 10078-10093.

[16] FEICHTENHOFER C, LI Y, HE K. Masked autoencoders as spatiotemporal learners[J]. Advances in neural information processing systems, 2022, 35: 35946-35958.

[17] WANG Y, LI K, LI Y, et al. Internvideo: General video foundation models via generative and discriminative learning[J]. arXiv preprint arXiv:2212.03191, 2022.

[18] GIRDHAR R, SINGH M, RAVI N, et al. Omnivore: A single model for many visual modalities[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16102-16112.

[19] GIRDHAR R, EL-NOUBY A, SINGH M, et al. OmniMAE: Single Model Masked Pretraining on Images and Videos." arXiv, Jun. 16, 2022. Accessed: Nov. 29, 2022[J].

[20] ADARSH P, RATHI P, KUMAR M. YOLO v3-Tiny: Object Detection and Recognition using one stage improved model[C]// 2020 6th international conference on advanced computing and communication systems (ICACCS). IEEE, 2020: 687-694.

[21] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[22] FANG H S, LI J, TANG H, et al. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[23] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[C]//European Conference on Computer Vision, Amsterdam Netherlands, 2016, 2011: 21-37

[24] JI S, XU W, YANG M, et al. 3D convolutional neural networks for heman action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221-231

[25] ZHI Y, HUI H. A fall detection method based on two-stream convolutional neural network[J]. Journal of Henan Normal University (Natural Science Edition), 2017, 45(3): 96-101

[26] KESKES O, NOUMEIR R. Vision-based fall detection using st-gcn[J]. IEEE Access, 2021, 9: 28224-28236.

备注：本文的彩色图表可从本刊官网（http//lyspkj.ijournal.cn）、中国知网、万方、维普、超星等数据库下载获取。